Apprentissage statistique en grande dimension et application au diagnostic oncologique par radiomique

Charles Bouveyron

Apprentissage statistique en grande dimension et application au diagnostic oncologique par radiomique

Statistical learning in high dimensions and application to cancer diagnostic with radiomics

Pr. Charles BOUVEYRON

Professeur de Mathématiques Appliquées, Chaire Inria en Science des Données
Laboratoire de Mathématiques J.A. Dieudonné, UMR CNRS 7351
Equipe-projet Epione, Inria Sophia-Antipolis
Université Côte d'Azur, Nice, France

Résumé : Avec l'augmentation des capacités de mesures, de nombreuses disciplines médicales ont vu leurs pratiques profondément modifiées du fait de la dimensionnalité des données acquises. Même si ces améliorations techniques font espérer des avancées importantes en recherche médicale, les méthodes d'apprentissage statistique mises en œuvre doivent être capables de faire face aux problèmes rencontrés dans les espaces de grande dimension. Les méthodes de classification dans des sous-espaces et les méthodes « sparses » introduites ces dernières années se proposent de répondre à cette attente. Cet article présente un rapide tour d'horizon de ces difficultés et des solutions proposés, ainsi qu'une illustration de l'usage d'une de ces solutions pour le diagnostic oncologique par radiomique.

Summary: With the increase in measurement capabilities, many medical disciplines have seen their practices deeply modified because of the dimensionality of the data. Although these technical improvements promise significant advances in medical research, the statistical learning methods must be able to cope with the problems encountered in those high-dimensional spaces. The subspace classification and "sparse" methods introduced in recent years propose to meet this expectation. This article presents a quick overview of these difficulties and the proposed solutions, as well as an illustration of the use of one of these solutions for oncology diagnosis with radiomics.

1. Introduction : apprentissage, fléau et bénédiction de la dimension

L'apprentissage statistique joue de nos jours un rôle croissant dans de nombreux domaines scientifiques aussi variés que la médecine, l'imagerie, la biologie ou l'astronomie. Les progrès scientifiques réalisés ces dernières années ont permis d'augmenter sensiblement les capacités de mesure et de calcul, et il est à présent difficile pour un opérateur humain de traiter de façon exhaustive ces données dans un temps raisonnable. En particulier, de nombreuses spécialités médicales, telles que l'imagerie médicale, la radiologie ou la génomique, ont bénéficié dans les dernières décennies d'évolutions importantes de leurs technologies. Dans certains cas, ces évolutions ont amené les spécialistes de ces domaines à devoir repenser leur pratique des données.

L'apprentissage statistique, qui doit être vu comme une sous-discipline de ce que l'on appelle aujourd'hui communément l'intelligence artificielle (IA), se propose alors de prendre le relais sur l'expert humain pour modéliser et synthétiser ces données complexes dans le but d'aider les praticiens à la prise de décision. Dans les applications médicales, la classification supervisée (ou analyse discriminante) est probablement la méthode d'apprentissage la plus utilisée pour le diagnostic ou le pronostic lié à des pathologies. Néanmoins, certaines situations pratiques correspondent à des problèmes théoriques qui ne sont pas entièrement résolus. Par exemple,

la classification de données de très grande dimension ou la classification de données corrélées sont des problèmes particulièrement présents en analyse d'images et biologie et pour lesquels les solutions actuelles, pourtant déjà très avancées, nécessitent que des recherches soient poursuivies.

En particulier, la grande dimension des données (nombre de variables important) pose un ensemble de problèmes à la statistique multivariée classique que l'on résume usuellement par le terme « fléau de la dimension ». Parmi les problèmes que posent la grande dimension des données on peut citer les problèmes numériques, les problèmes d'inférence ou les problèmes de biais des estimateurs. Il a donc été nécessaire de développer ces dernières années des méthodes capables de pallier ces problèmes. Nous allons dans cet article explorer rapidement les problèmes et les espoirs dus à la grande dimension des données, ainsi que quelques méthodes récentes d'apprentissage statistique qui permettent une meilleur analyse et compréhension des données médicales.

1.1. Fléau et bénédiction de la dimension

S'il est une expression qui est classiquement associée aux données de grande dimension, c'est certainement celle de « fléau de la dimension », qui fut introduite par R. Bellman à la fin des années 50. R. Bellman a utilisé cette locution pour la première fois dans la préface de son livre « *Dynamic programming* » pour résumer l'ensemble des difficultés posées par les espaces de grande dimension. En particulier, les espaces de grande dimension ont des propriétés souvent surprenantes et il est en général difficile d'extrapoler les propriétés de ce qui est connu des espaces à deux ou trois dimensions aux espaces de plus grande dimension. Une façon simple d'observer cela est de s'intéresser au volume de l'hyper-sphère[1] unité en fonction de la dimension de l'espace. Ce volume est donné par la formule $V(p) = \pi^{p/2}/\Gamma(p/2+1)$ et la Figure 1 permet de visualiser l'évolution de cette fonction en regard de la dimension p de l'espace. Il apparait clairement que, au-delà de la dimension 5, le volume de l'hyper-sphère tend vers zéro très rapidement, ce qui est clairement contre-intuitif. Du point de vue de la statistique, l'apprentissage dans de tels espaces se heurte à des problèmes numériques liés à la sur-paramétrisation[2] ou des problèmes de biais d'estimation.

En revanche, les espaces de grande dimension présentent également des caractéristiques qui apportent un peu d'espoir, et qui peuvent donc être vues comme une « bénédiction », en regard des fléaux suscités. Le phénomène de « l'espace vide », mis en évidence à la fin des années 70, traduit le fait que les espaces de grande dimension sont principalement vides et que les données de grande dimension se regroupent dans des sous-espaces de dimension faibles. Cela peut clairement être un avantage quand on souhaite discriminer des classes d'individus, mais pour cela il faut être en capacité de pallier les problèmes numériques et biais d'estimation. La Figure 2 montre en effet qu'un classifieur « oracle » profite de ce phénomène alors que le classifieur appris à partir des données dégrade la performance qu'il obtient en faible dimension. L'espoir d'exploiter le phénomène de l'espace vide en classification dépend donc de notre capacité à résoudre efficacement le problème d'inférence statistique.

1.2. Approches usuelles pour l'apprentissage en grande dimension

Les approches usuelles pour contourner les problèmes posés par la grande dimension des données sont la réduction de dimension, la régularisation et l'usage de modèles parcimonieux. La réduction de dimension est certainement l'approche la plus ancienne et la plus utilisée en pratique. C'est en effet la manière la plus directe de contrer le problème de la grande

[1] L'hyper-sphère est la généralisation de la sphère usuelle dans des espaces à plus que 3 dimensions.

[2] Certaines méthodes d'apprentissage ont un nombre de paramètres qui croît avec la dimension de l'espace, ce qui peut poser problème en grande dimension.

dimensionnalité mais qui a le désavantage de potentiellement engendrer une perte d'information discriminante. La régularisation s'attaque quant à elle aux problèmes numériques, notamment dus à la grande colinéarité[3] des variables. Ces techniques de régularisation peuvent s'avérer en revanche difficile à paramétrer. Enfin, les approches parcimonieuses contraignent la modélisation pour réduire leur niveau de paramétrisation, quitte à faire parfois des hypothèses fortes telles que l'indépendance conditionnelle des variables[4].

2. Avancées récentes de l'apprentissage en grande dimension

Les méthodes récentes de classification de données de grande dimension exploitent quant à elles pleinement le phénomène de l'espace vide. D'une part, parmi les méthodes discriminatives (*i.e.* donc le but unique est de construire une fonction de classification), les méthodes à noyaux (*support vector machines*) et les réseaux de neurones (*convolutional neural networks*) n'hésitent pas à projeter les données dans des espaces de dimension grande, voir infinie, à l'aide d'un projecteur non-linéaire, pour faciliter la séparation des classes. D'autre part, parmi les méthodes génératives (*i.e.* qui modélisent les classes et déduisent de cette modélisation une règle de classification), les méthodes de sous-espaces ou de sélection de variables sont celles qui exploitent le mieux les qualités des espaces de grande dimension. La différence entre méthodes discriminatives et génératives est illustrée dans la Figure 2.

2.1. Classification dans des sous-espaces

Le travail le plus ancien dans ce contexte est le travail séminal de R. Fisher qui introduisit en 1936 l'analyse discriminante linéaire (LDA en anglais et qui deviendra connue comme l'analyse discriminante de Fisher). L'objectif de LDA est de trouver un sous-espace de faible dimension qui discrimine au mieux les classes. Même si LDA peut souffrir de la colinéarité des variables, elle reste une méthode étalon qui fournit le plus souvent des résultats très satisfaisants. Parmi les méthodes les plus récentes et couramment utilisées, PLSDA et HDDA font toutes les deux l'hypothèse que les données vivent dans des sous-espaces. PLSDA (*partial least square discriminant analysis*, Barker et Rayens (2003)) recherche des représentations latentes des données et de la variable à prédire telles que la covariance entre les deux soit maximale. PSLDA est une méthode très utilisée dans les disciplines telles que la génomique ou la métabolomique. HDDA (*high-dimensional discriminant analysis*, Bouveyron *et al.* (2007)) s'appuie sur une modélisation statistique qui suppose que les données de chaque classe vivent dans des sous-espaces différents et de dimensions intrinsèques différentes. Ainsi HDDA peut modéliser de façon très fines les données et ainsi fournir un classifieur très performant. La Figure 3 illustre le principe de modélisation dans des sous-espaces spécifiques aux classes. Le lecteur souhaitant plus de détails sur ce type d'approches peut consulter Bouveyron et Brunet-Saumard (2014).

2.2. Classification par sélection de variables

La sélection de variable a également été exploré très tôt pour contourner le problème de la grande dimension tout en conservant l'avantage de l'interprétation vis-à-vis des variables d'origines. Le critère de Fisher peut être utilisé dans ce cadre également et le Lambda de Wilks permet notamment de construire des tests statistiques pour décider de l'utilité d'une variable. Malheureusement, ces approches se sont rapidement heurtées à la combinatoire de l'exploration des sous-ensembles de nombreuses variables. Il a fallu attendre le début des

[3] Deux vecteurs u et v sont dits colinéaires s'il existe un scalaire λ tel que $v = \lambda u$.
[4] Certains modèles supposent que la matrice de covariance de chacune des classes est diagonale.

années 2000 pour voir émerger une nouvelle approche : la sélection de variable par « sparsité Lasso ». Les approches Lasso opèrent une sélection de variables en ajoutant à la fonction objective d'apprentissage une pénalisation de type ℓ_1. L'approche Lasso nécessite en revanche de définir le niveau de parcimonie et cela est en général fait par validation-croisée. Parmi les méthodes dites « sparses » de classification, on peut citer *sparse discriminant analysis* (SDA), proposée par Witten et Tibshirani (2011), qui introduit de la parcimonie au sein de la méthode LDA par pénalisation ℓ_1.

2.3. Sélection de variables en classification par parcimonie bayésienne

Tout récemment, Mattei *et al.* (2016) ont introduit dans le cadre de l'analyse en composantes principales (ACP) une parcimonie structurée induite par un a priori bayésien[5]. La méthode résultante, nommée gsPPCA, permet alors de réduire la dimensionnalité d'un jeu de données tout en sélectionnant les variables utiles à la bonne description des données. Cette approche a également été étendue au cadre de la classification en reportant la parcimonie structurée par a priori bayésien au sein d'HDDA. Ainsi, la méthode résultante *sparse* HDDA (sHDDA, Orlhac *et al.* (2018)) permet de sélectionner les variables nécessaires à la modélisation de chacune des classes indépendamment. Pour ce faire, sHDDA suppose que chaque classe est distribuée selon une loi normale :

$$X|Z = k \sim \mathcal{N}(\mu_k, \Sigma_k), \text{où } \Sigma_k = V_k S_k V_k^t + b_k I_p.$$

avec $V_k = \text{diag}(v_1, ..., v_k, ..., v_p)$ une matrice diagonale binaire indiquant si chacune des p variables est utile ou non pour modéliser la classe k. Il est à noter que, contrairement aux approches Lasso, gsPPCA et sHDDA ne recourent pas à la validation croisée pour déterminer le nombre de variables pertinentes car ce paramètre est inhérent à la modélisation. En effet, gsPPCA et sHDDA optimisent la vraisemblance marginale par rapport aux variables v_k sur un chemin de modèles et le modèle le plus vraisemblable est retenu. La Figure 4 illustre le choix du nombre de variables pertinentes pour une des classes.

3. Application au diagnostic oncologique par radiomique

La radiomique est une technique émergente en recherche médicale qui consiste en l'extraction de caractéristiques tumorales à partir d'images médicales, telles que l'IRM, CT scan ou PET scan. Les caractéristiques extraites sur les tumeurs décrivent leur hétérogénéité, leur forme et leur texture. Le nombre de variables extraites peut varier, selon la technologie utilisée, de quelques dizaines à plusieurs centaines de variables. L'apprentissage d'un classifieur performant est complexifié par le fait que la plupart des études incluent moins d'une centaine de patients. Le ratio nombre de patients / nombre de variables est alors très défavorable à l'estimation statistique. Les méthodes de classification telles que sHDDA permettent d'espérer discriminer efficacement le sous-type de lésion (ce qui pourrait diminuer le nombre de biopsies nécessaires) tout en sélectionnant les variables pertinentes pour décrire chaque sous-type (ce qui apporterait une meilleure compréhension des différents sous-types).

Une étude très récente de Orlhac *et al.* (2018b) illustre les possibilités de la radiomique pour la prédiction du sous-type histologique en cancer du sein. Dans cette étude, les auteurs disposent d'une cohorte de 26 patientes atteintes d'un cancer du sein et traitées au Centre Antoine Lacassagne de Nice. Parmi ces 26 patientes, 7 avaient une lésion de type triple-négatif, qui est un type de tumeurs particulièrement agressif. Pour l'ensemble des patientes,

[5] Certains « paramètres » du modèle sont alors vus comme des variables aléatoires ayant leur propre distribution a priori.

des images TEP sont disponibles et ont permis d'extraire 43 variables radiomiques en utilisant le logiciel LIFEx. En outre, une analyse de la pièce opératoire avec un spectromètre de masse a permis de quantifier 1500 métabolites identifiés dans la base de données Human Metabolome. Après avoir montré qu'une grande partie des 43 variables radiomiques avait une forte corrélation avec au moins 50 métabolites, les auteurs ont comparé 5 méthodes de classification en grande dimension pour la prédiction du sous-type histologique (triple-négatif contre le reste), et ce sur les variables radiomiques puis sur les variables métabolomiques. La Figure 5 présente les résultats de classification (en validation croisée « 25-folds »). Le score de Youden (Y=sensibilité+spécificité-1) est utilisé pour évaluer la performance. Rappelons qu'un classifieur parfait aura un Youden égal à 1. On remarque que les méthodes HDDA et sHDDA (qui opère une sélection de variables) sont particulièrement performantes et que les meilleurs résultats sont obtenus à partir des variables radiomiques. La Figure 6 permet en outre d'observer la sélection de variables proposée par sHDDA, en fonction du nombre de réplications. On notera que la sélection est particulièrement stable, et ce malgré le faible volume de données.

4. Conclusion

Dans ce court article, nous avons présenté un tour d'horizon des problématiques et solutions existantes en classification en grande dimension. Les recherches menées en apprentissage statistique ces quinze dernières années ont permis des avancées importantes, qui permettent aujourd'hui de faire face à des tâches de classification particulièrement ardues. Comme nous l'avons illustré, ces outils avancés peuvent être mis en œuvre afin d'obtenir des résultats significatifs dans des applications médicales comme l'oncologie sur la base de technologies récentes telles que la radiomique ou la métabolomique. Cette chaîne de traitement de données pourrait permettre à moyen terme de prédire le sous-type histologique de la lésion sans avoir recourt à une biopsie. Un objectif à plus long terme serait de pouvoir s'appuyer sur ces technologies pour prédire la réponse au traitement dès les premières consultations et ainsi éviter d'attendre un ou plusieurs cycles de traitement pour observer cette réponse. Les méthodes d'apprentissage statistique ont donc un rôle important à jouer dans l'exploitation des données « omics » pour leur usage en médecine.

Références

M. Barker and W. Rayens, *Partial least squares for discrimination*, Journal of Chemometrics, vol. 17(3), pp. 166–173, 2003.

C. Bouveyron, S. Girard and C. Schmid, *High-Dimensional Data Clustering*, Computational Statistics and Data Analysis, vol. 52 (1), pp. 502-519, 2007.

C. Bouveyron and C. Brunet-Saumard, *Model-based clustering of high-dimensional data: A review,* Computational Statistics and Data Analysis, vol. 71, pp. 52–78, 2013.

C. Bouveyron, P. Latouche and P.-A. Mattei, *Bayesian Variable Selection for Globally Sparse Probabilistic PCA*, Preprint HAL n°01310409, Université Paris Descartes, 2016.

R.A. Fisher, *The use of multiple measurements in taxonomic problems*, The Annals of Eugenics, vol. 7, pp. 179–188, 1936.

F. Orlhac, P.-A. Mattei, C. Bouveyron and N. Ayache, *Class-specific Variable Selection in High-Dimensional Discriminant Analysis through Bayesian Sparsity*, Preprint HAL n°xxxxxxxx, Université Côte d'Azur, 2018.

F. Orlhac, O. Humbert, T. Pourcher, L. Jing, J.-M. Guigonis, J. Darcourt, C. Bouveyron, N. Ayache, *Statistical analysis of PET radiomic features and metabolomic data: prediction of triple-negative breast cancer*, SNMMI 2018 annual meeting, Journal of Nuclear Medicine, vol. 59, p. 1755, 2018.
D. Witten and R. Tibshirani, *Penalized classification using Fisher's linear discriminant*, Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 73(5), pp. 753–772, 2011.

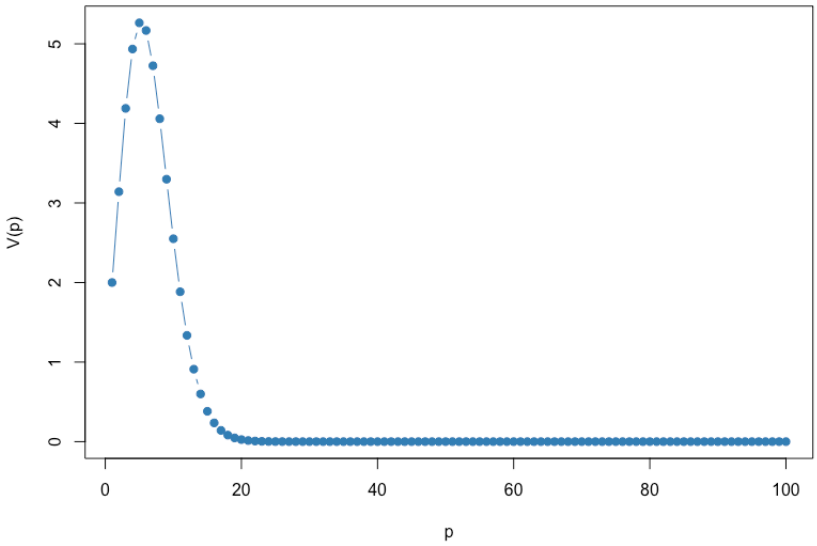

Figure 1. Evolution du volume de l'hyper-sphère unité en fonction de la dimension de l'espace.

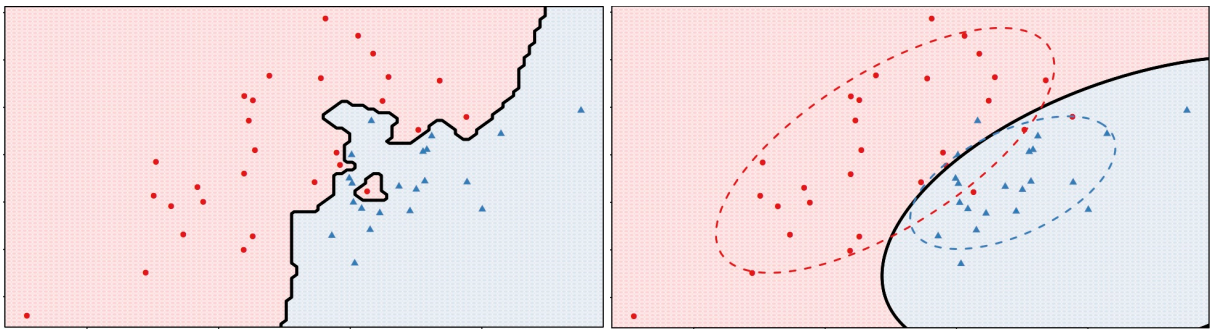

Figure 2. Frontières de classification d'un classifieur discriminatif (gauche) et d'un classifieur génératif (droite).

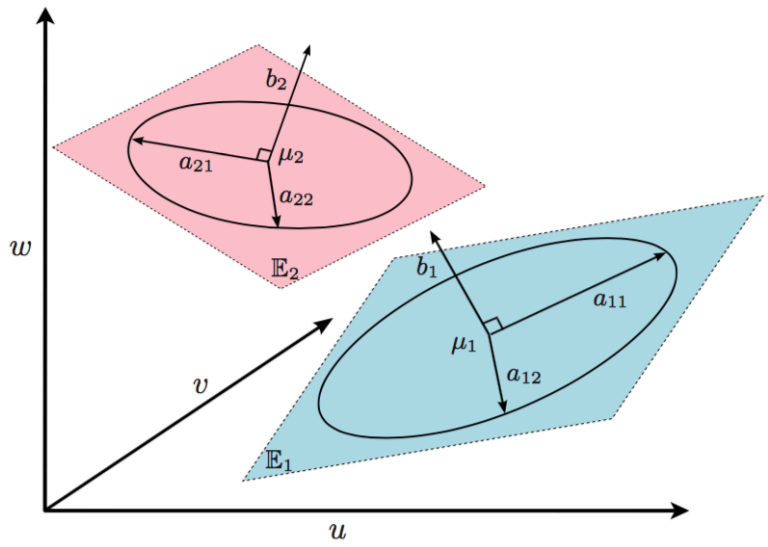

Figure 3. Illustration de la modélisation des classes dans des sous-espaces spécifiques.

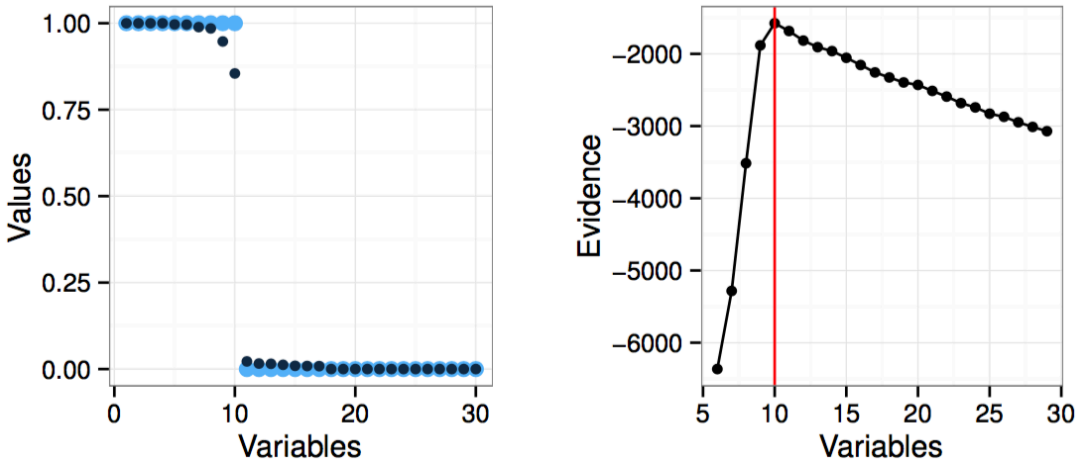

Figure 4. Illustration de la sélection de variable opérée par sHDDA pour une classe.

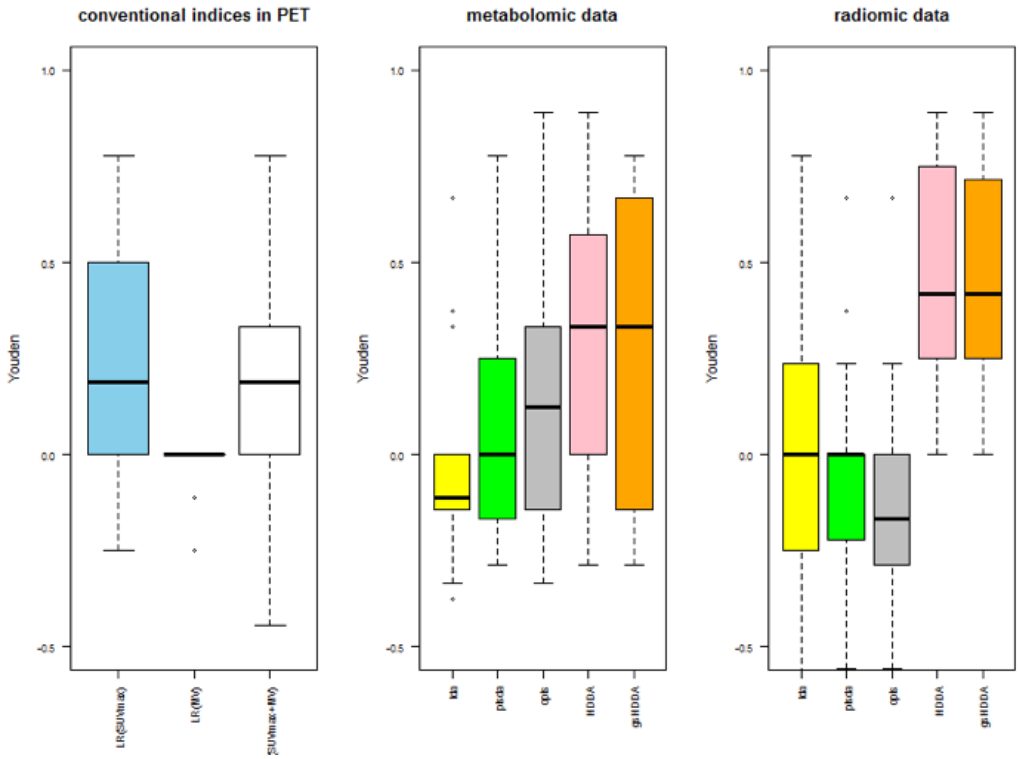

Figure 5. Performance de classification (score de Youden) pour la prédiction du type triple-négatif avec différentes méthodes de classification (avec l'aimable autorisation de Orlhac et al. (2018)).

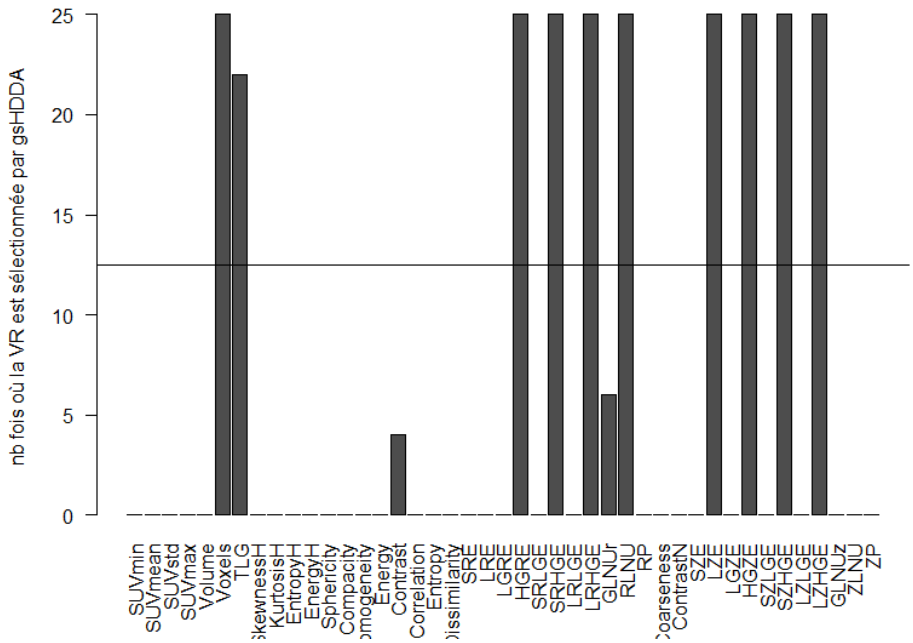

Figure 6. Variables radiomiques sélectionnées par la méthode sHDDA pour la prédiction du type triple-négatif (avec l'aimable autorisation de Orlhac et al. (2018)).